1/5/1 (Item 1 from file: 351)
DIALOG(R)File 351:Derwent WPI

012677564 **Image available**
WPI Acc No: 1999-483671/ **199941**
XRPX Acc No: N99-360701

Data forwarding procedure in multiprocessor system - involves resending page where failure occurs during address conversion unit, by referring two failure page tables

Patent Assignee: NEC CORP (NIDE)
Number of Countries: 001 Number of Patents: 002
Patent Family:

Patent No	Kind	Date	Applicat No	Kind	Date	Week	
JP 11203260	A	19990730	JP 986578	A	19980116	199941	B
JP 3237599	B2	20011210	JP 986578	A	19980116	200203	

Priority Applications (No Type Date): JP 986578 A 19980116
Patent Details:

Patent No	Kind	Lan	Pg	Main IPC	Filing Notes
JP 11203260	A		16	G06F-015/163	
JP 3237599	B2		16	G06F-015/177	Previous Publ. patent JP 11203260

Abstract (Basic): JP 11203260 A
NOVELTY - An address converter performs conversion between physical address and logic address. A transmitter (171) transmits data on logic address to other nodes via communication channel (27). A memory records page where failure occurs, during address conversion. A resending unit resends stored page after confirming existence of address conversion failure with reference to failure page tables (341,641).
USE - In multiprocessor system.
ADVANTAGE - Data forwarding between nodes is done at high speed.
DESCRIPTION OF DRAWING(S) - The figure shows functional block diagram of function component of multiprocessor system. (27) Communication channel; (171) Transmitter; (341,641) Failure page tables.
Dwg.1/13

Title Terms: DATA; FORWARDING; PROCEDURE; MULTIPROCESSOR; SYSTEM; PAGE; FAIL; OCCUR; ADDRESS; CONVERT; UNIT; REFER; TWO; FAIL; PAGE; TABLE
Derwent Class: T01
International Patent Class (Main): G06F-015/163; G06F-015/177
International Patent Class (Additional): G06F-012/10
File Segment: EPI

1/5/2 (Item 1 from file: 347)
DIALOG(R)File 347:JAPIO

06261680 **Image available**
MULTIPROCESSOR SYSTEM AND DATA TRANSFER METHOD IN MULTIPROCESSOR SYSTEM

PUB. NO.: 11-203260 A]
PUBLISHED: July 30, 1999 (19990730)
INVENTOR(s): SUGAWARA TOMOYOSHI
APPLICANT(s): NEC CORP
APPL. NO.: 10-006578 [JP 986578]
FILED: January 16, 1998 (19980116)
INTL CLASS: G06F-015/163

ABSTRACT

PROBLEM TO BE SOLVED: To perform data transfer between plural nodes at high speed.

SOLUTION: When a page-out occurs to data to be transferred and a

transmission part 171 or a reception part 271 fails address conversion during data transmission by a communication hardware 27, the transmission part 171 or the reception part 271 demands interruption. An interruption processing 362 records a page which fails in the address translation in a transmission failure page table 341 or a reception failure page table 641, and skips the page to continue data transfer. After a series of data transfer is completed, a process 11 having requested data transmission refers to both of the failure page tables and examines whether or not there is a failure in the address translation. When there is a failure in the address conversion, the process 11 loads the page in a memory 13 and retransmits it.

(19)日本国特許庁（ＪＰ）　(12)特　許　公　報（Ｂ２）　(11)特許番号
特許第3237599号
(P3237599)

(45)発行日　平成13年12月10日(2001.12.10)　(24)登録日　平成13年10月５日(2001.10.5)

(51)Int.Cl.'	識別記号	ＦＩ	
Ｇ０６Ｆ　15/177	６７６	Ｇ０６Ｆ　15/177	６７６Ａ
12/10	５０７	12/10	５０７Ｂ

請求項の数６(全 16 頁)

(21)出願番号　特願平10－6578

(22)出願日　平成10年１月16日(1998.1.16)

(65)公開番号　特開平11－203260
(43)公開日　平成11年７月30日(1999.7.30)
審査請求日　平成10年１月16日(1998.1.16)

前置審査

(73)特許権者　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(72)発明者　菅原　智義
東京都港区芝五丁目７番１号　日本電気株式会社内
(74)代理人　100082935
弁理士　京本　直樹　(外２名)

審査官　久保　正典

最終頁に続く

(54)【発明の名称】　マルチプロセッサシステム及びマルチプロセッサシステムにおけるデータ転送方法

(57)【特許請求の範囲】
【請求項１】互いに通信路を介して接続された複数のノードから構成されるマルチプロセッサシステムであって、
前記複数のノードのうちの少なくとも１つは、
物理アドレス空間を提供する第１の記憶手段と、
論理アドレス空間を提供する第２の記憶手段と、
所定のアドレス変換単位毎に物理アドレスと論理アドレスとの間のアドレス変換を行う第１のアドレス変換手段と、
前記論理アドレス空間に存在するデータを前記通信路を介して他のノードに送信する第１の送信手段と、
前記第１の送信手段によって送信される前記データのうち前記第１の記憶手段が提供する前記物理アドレス空間に存在しなかったアドレス変換単位に関する情報を記録

する第１の変換単位情報記録手段と、
前記第１の送信手段による前記データの送信が終了した後、前記第１の変換単位情報記録手段に記録されている情報に対応するアドレス変換単位のデータを前記通信路を介して前記他のノードに再送する再送手段と、を備え、
前記複数のノードのうちの他の少なくとも１つは、
物理アドレス空間を提供する第３の記憶手段と、
論理アドレス空間を提供する第４の記憶手段と、
所定のアドレス変換単位毎に物理アドレスと論理アドレスとの間のアドレス変換を行う第２のアドレス変換手段と、
前記複数のノードのうちの少なくとも１つから送信されたデータを受信する受信手段と、
前記受信手段が受信したデータを記憶すべき論理アドレ

スに対応する物理アドレスが前記第３の記憶手段に存在しないアドレス変換単位に関する情報を記録する第２の変換単位情報記録手段と、を備え、
前記複数のノードのうちの少なくとも１つは、
前記通信路を介して前記第２の変換単位情報記録手段に記録された情報を読み出す変換単位情報読み出し手段と、
前記変換単位情報読み出し手段が読み出した情報に対応するデータを前記第２のアドレス変換手段にアドレス変換させて前記第４の記憶手段から前記第３の記憶手段に書き込む手段とを備え、
前記再送手段は、前記変換単位情報読み出し手段が読み出した情報に対応するアドレス変換単位のデータをさらに前記複数のノードのうちの少なくとも１つから前記通信路を介して前記複数のノードのうちの他の少なくとも１つに再送することを特徴とするマルチプロセッサシステム。

【請求項２】前記受信手段は、受信したデータのうち前記第３の記憶手段が提供する前記物理アドレス空間に存在しなかったアドレス変換単位のデータを破棄することを特徴とする請求項１に記載のマルチプロセッサシステム。

【請求項３】前記送信されるデータのうちの最後の送信単位には、該送信単位が最後のものであることを示す情報が含まれ、
前記複数のノードのうちの他の少なくとも１つは、
前記受信手段が前記送信単位のうちの最後のデータを受信したときに、前記前記第２の変換単位情報記録手段に記録されている情報を参照する第２の変換単位情報参照手段と、
前記第２の変換単位情報参照手段による参照の結果、いずれかのアドレス変換単位を記憶すべき論理アドレスに対応する物理アドレスが前記第３の記憶手段に存在しなかったことを示すときに、前記最後の送信単位を含むアドレス変換単位に関する情報を前記第２の変換単位情報記録手段に記録させる第２の変換単位情報制御手段と、
前記第２の変換単位情報参照手段による参照の結果、いずれかのアドレス変換単位を記憶すべき論理アドレスに対応する物理アドレスが前記第３の記憶手段に存在しなかったことを示すときに、前記最後の送信単位のデータを破棄する破棄手段とをさらに備えることを特徴とする請求項１または２に記載のマルチプロセッサシステム。

【請求項４】前記複数のノードのうちのさらに他の少なくとも１つは、
前記第１の記憶手段が提供する物理アドレス空間に対する論理アドレス空間をさらに提供する第５の記憶手段と、
前記複数のノードのうちの少なくとも１つからの要求に従うアドレス変換単位を送信する第２の送信手段とを備え、

前記複数のノードのうちの少なくとも１つは、
前記第１の送信手段による前記データの送信が終了した後、前記第１の変換単位情報記録手段に記録されている情報に対応するアドレス変換単位が前記第５の記憶手段に記憶されているかどうかを判定する判定手段と、
前記判定手段による判定の結果、前記第５の記憶手段に記憶されていると判定されたアドレス変換単位を前記第２の送信手段に送信させるべき前記要求を前記通信路を介して行う送信要求手段と、をさらに備えることを特徴とする請求項１乃至３のいずれか１項に記載のマルチプロセッサシステム。

【請求項５】前記再送手段は、
前記第１の変換単位情報記録手段に記録されている情報に対応するデータを前記第１のアドレス変換手段にアドレス変換させて前記第２の記憶手段から前記第１の記憶手段に書き込む手段を備え、
該手段によって前記第１の記憶手段に書き込まれたデータを前記通信路を介して前記他のノードに再送することを特徴とする請求項１乃至４のいずれか１項に記載のマルチプロセッサシステム。

【請求項６】互いに通信路を介して接続され、それぞれ物理アドレス空間を提供する第１の記憶装置と、論理アドレス空間を提供する第２の記憶装置と、所定のアドレス変換単位毎に物理アドレスと論理アドレスとの間のアドレス変換を行うアドレス変換機構とを有する複数のノードから構成されるマルチプロセッサシステムにおけるデータ転送方法であって、
前記複数のノードのうちの少なくとも１つから、該ノードの論理アドレス空間に存在するデータを前記通信路を介して他のノードに送信する送信ステップと、
前記送信ステップで送信される前記データのうち前記第１の記憶装置が提供する前記物理アドレス空間に存在しなかったアドレス変換単位に関する情報を記録する第１の変換単位情報記録ステップと、
前記送信ステップでの前記データの送信が終了した後、前記第１の変換単位情報記録ステップで記録した情報に対応するアドレス変換単位のデータを前記複数のノードのうちの少なくとも１つから前記通信路を介して前記他のノードに再送する再送ステップと、
前記送信ステップで前記複数のノードの少なくとも１つから送信されたデータを受信した前記他のノードが、該データを記憶すべき論理アドレスに対応する物理アドレスが前記第１の記憶装置に存在しないアドレス変換単位に関する情報を記録する第２の変換単位情報記録ステップとを含み、
前記再送ステップは、前記第２の変換単位情報記録ステップで記録した情報に対応するアドレス変換単位のデータをさらに前記複数のノードのうちの少なくとも１つから前記通信路を介して前記他のノードに再送することを特徴とするマルチプロセッサシステムにおけるデータ転

送方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、マルチプロセッサシステム及びマルチプロセッサシステムにおけるデータ転送方法に関し、特に各ノードの通信ハードウェアが仮想アドレスによるデータ転送機能をサポートするマルチプロセッサシステムに適用するものに関する。

【０００２】

【従来の技術】従来、ネットワークを介して複数のノードが結合されたマルチプロセッサシステムにおいて、ノード間でデータ転送を行う場合には、オペレーティングシステム（ＯＳ）の機能を利用する方法が一般的であった。しかしながら、この方法では、ＯＳをシステムコールするための通信遅延時間が大きくなり、また、ユーザ空間とシステム空間との間でデータをコピーしなければならないため、データの転送速度が遅くなるという問題があった。

【０００３】そこで、近年の分散メモリ型並列マシンやクラスタ型マルチプロセッサシステムでは、ノード間通信の機能として仮想アドレスによるデータ転送をサポートするものも用いられている。このようなシステムでは、ユーザプロセスはＯＳを介することなく通信ハードウェアによるデータ転送を可能としている。しかしながら、このようなシステムでは、送信されるデータや受信したデータを格納すべきバッファは、主記憶装置上に存在する保証がなく、アドレス変換に失敗して通信ハードウェアの動作が停止する可能性があった。

【０００４】この問題を解決するため、特開平６－１９８５６号公報において、予め通信領域を特定し、その部分はページアウトされないようにＯＳに登録しておくことによって、データの転送時にアドレス変換の失敗が生じないようにする方法（第１の方法）が提案されている。

【０００５】また、アドレス変換の失敗があった時点でデータの送信あるいは受信を停止して、通信ハードウェアから割り込みをかけてオペレーティングシステムを呼び出し、データの送信或いは受信を一旦停止して、ページフォルトに対する処理を行ってからデータの送信或いは受信を再開する方法（第２の方法）も提案されている。

【０００６】

【発明が解決しようとする課題】しかしながら、上記第１の方法では、ユーザプロセス中の任意のデータを転送するためには、結局データを通信領域にコピーする必要があり、データ転送の速度を低下させるという問題があった。また、上記第２の方法では、ページフォルトの処理の間、データの送信或いは受信が待たされることとなるので、すべてのデータについての転送を終了するまでに時間がかかるという問題があった。

【０００７】本発明は、上記従来例の問題点を解消するためになされたものであり、複数のノード間でのデータ転送を高速に行うことができるマルチプロセッサシステム及びこのようなマルチプロセッサシステムにおけるデータ転送方法を提供することを目的とする。

【０００８】

【課題を解決するための手段】上記目的を達成するため、本発明の第１の観点にかかるマルチプロセッサシステムは、互いに通信路を介して接続された複数のノードから構成されるマルチプロセッサシステムであって、前記複数のノードのうちの少なくとも１つは、物理アドレス空間を提供する第１の記憶手段と、論理アドレス空間を提供する第２の記憶手段と、所定のアドレス変換単位毎に物理アドレスと論理アドレスとの間のアドレス変換を行う第１のアドレス変換手段と、前記論理アドレス空間に存在するデータを前記通信路を介して他のノードに送信する第１の送信手段と、前記第１の送信手段によって送信される前記データのうち前記第１の記憶手段が提供する前記物理アドレス空間に存在しなかったアドレス変換単位に関する情報を記録する第１の変換単位情報記録手段と、前記第１の送信手段による前記データの送信が終了した後、前記第１の変換単位情報記録手段に記録されている情報に対応するアドレス変換単位のデータを前記通信路を介して前記他のノードに再送する再送手段と、を備え、前記送信されるデータのうちの最後の送信単位には、該送信単位が最後のものであることを示す情報が含まれ、前記複数のノードのうちの少なくとも１つは、前記第１の送信手段が前記送信単位のうちの最後のデータを送信するときに、前記第１の変換単位情報記録手段に記録されている情報を参照する第１の変換単位情報参照手段と、前記第１の変換単位情報参照手段による参照の結果、いずれかのアドレス変換単位が前記第１の記憶手段が提供する前記物理アドレス空間に存在しなかったことを示すときに、前記最後の送信単位を含むアドレス変換単位に関する情報を前記第１の変換単位情報記録手段に記録させる第１の変換単位情報制御手段と、前記第１の変換単位情報参照手段による参照の結果、いずれかのアドレス変換単位が前記第１の記憶手段が提供する前記物理アドレス空間に存在しなかったことを示すときに、前記第１の送信手段による最後の送信単位のデータの送信を停止させる送信停止手段とをさらに備えることを特徴とする。

【０００９】

【００１０】

【００１１】上記マルチプロセッサシステムでは、前記送信手段によって送信されるデータのうちで物理アドレス空間に存在しないアドレス変換単位、例えば、前記第２の記憶手段にページアウトされているページは、それに関する情報（例えば、ページ番号）が前記第１の変換情報記録手段に記録された後、前記再送手段によって再

送される。このため、前記変換情報記録手段への記録の際にページイン要求を出し、それと並行してデータ処理を続行させること（プリページング）が可能となる。これにより、ページイン処理とデータ転送とをオーバーラップさせることができ、全体としてデータ転送を高速化することができる。

【００１２】また、前記再送手段は、前記第１の変換単位情報記録手段に記録されている情報に従ってアドレス変換単位を再送すればよい。また、前記第１のアドレス変換手段は、前記第１の変換単位情報記録手段に記録されている情報に従ってアドレス変換単位のアドレス変換を行えばよい。このため、必要のないアドレス変換やデータの再送をしなくてもよく、無駄を防ぐことができる。そして、最後の送信単位は、前記再送手段による再送があっても、全体のデータの中で必ず最後に送信されることとなるので、例えば、データ中の最後の１ビットを受信完了を示すフラグとして用いるプログラムでも、正しく動作させることが可能となる。

【００１３】

【００１４】

【００１５】上記目的を達成するため、本発明の第２の観点にかかるマルチプロセッサシステムは、互いに通信路を介して接続された複数のノードから構成されるマルチプロセッサシステムであって、前記複数のノードのうちの少なくとも１つは、物理アドレス空間を提供する第１の記憶手段と、論理アドレス空間を提供する第２の記憶手段と、所定のアドレス変換単位毎に物理アドレスと論理アドレスとの間のアドレス変換を行う第１のアドレス変換手段と、前記論理アドレス空間に存在するデータを前記通信路を介して他のノードに送信する第１の送信手段と、前記第１の送信手段によって送信される前記データのうち前記第１の記憶手段が提供する前記物理アドレス空間に存在しなかったアドレス変換単位に関する情報を記録する第１の変換単位情報記録手段と、前記第１の送信手段による前記データの送信が終了した後、前記第１の変換単位情報記録手段に記録されている情報に対応するアドレス変換単位のデータを前記通信路を介して前記他のノードに再送する再送手段と、を備え、前記複数のノードのうちの他の少なくとも１つは、物理アドレス空間を提供する第３の記憶手段と、論理アドレス空間を提供する第４の記憶手段と、所定のアドレス変換単位毎に物理アドレスと論理アドレスとの間のアドレス変換を行う第２のアドレス変換手段と、前記複数のノードのうちの少なくとも１つから送信されたデータを受信する受信手段と、前記受信手段が受信したデータを記憶すべき論理アドレスに対応する物理アドレスが前記第３の記憶手段に存在しないアドレス変換単位に関する情報を記録する第２の変換単位情報記録手段と、を備え、前記複数のノードのうちの少なくとも１つは、前記通信路を介して前記第２の変換単位情報記録手段に記録された情報

を読み出す変換単位情報読み出し手段と、前記変換単位情報読み出し手段が読み出した情報に対応するデータを前記第２のアドレス変換手段にアドレス変換させて前記第４の記憶手段から前記第３の記憶手段に書き込む手段とを備え、前記再送手段は、前記変換単位情報読み出し手段が読み出した情報に対応するアドレス変換単位のデータをさらに前記複数のノードのうちの少なくとも１つから前記通信路を介して前記複数のノードのうちの他の少なくとも１つに再送することを特徴とする。

【００１６】このように構成した場合、上記マルチプロセッサシステムにおいて、前記受信手段は、受信したデータのうち前記第３の記憶手段が提供する前記物理アドレス空間に存在しなかったアドレス変換単位のデータを破棄するものとすることができる。

【００１７】この場合、受信側となるノードで受信したアドレス変換単位を記憶すべき物理アドレスが前記第３の記憶手段にないとき、例えば、受信したデータを記憶すべきページのうち前記第４の記憶手段にページアウトされているページは、それに関する情報（例えば、ページ番号）が前記第２の変換情報記録手段に記録された後、前記再送手段によって再送される。このため、前記変換情報記録手段への記録の際にページイン要求を出し、それと並行してデータ処理を続行させること（プリページング）が可能となる。これにより、ページイン処理とデータ転送とをオーバーラップさせることができ、全体としてデータ転送を高速化することができる。

【００１８】このように構成した場合にも、また、前記送信されるデータのうちの最後の送信単位には、該送信単位が最後のものであることを示す情報が含まれ、前記複数のノードのうちの他の少なくとも１つは、前記受信手段が前記送信単位のうちの最後のデータを受信したときに、前記前記第２の変換単位情報記録手段に記録されている情報を参照する第２の変換単位情報参照手段と、前記第２の変換単位情報参照手段による参照の結果、いずれかのアドレス変換単位を記憶すべき論理アドレスに対応する物理アドレスが前記第３の記憶手段に存在しなかったことを示すときに、前記最後の送信単位を含むアドレス変換単位に関する情報を前記第２の変換単位情報記録手段に記録させる第２の変換単位情報制御手段と、前記第２の変換単位情報参照手段による参照の結果、いずれかのアドレス変換単位を記憶すべき論理アドレスに対応する物理アドレスが前記第３の記憶手段に存在しなかったことを示すときに、前記最後の送信単位のデータを破棄する破棄手段とをさらに備えるものとすることができる。

【００１９】この場合、最後の送信単位は、前記再送手段による再送があっても、全体のデータの中で必ず最後に送信されることとなるので、例えば、データ中の最後の１ビットを受信完了を示すフラグとして用いるプログラムでも、正しく動作させることが可能となる。

【００２０】上記第１、第２の観点にかかるマルチプロセッサシステムにおいて、前記複数のノードのうちのさらに他の少なくとも１つは、前記第１の記憶手段が提供する物理アドレス空間に対する論理アドレス空間をさらに提供する第５の記憶手段と、前記複数のノードのうちの少なくとも１つからの要求に従うアドレス変換単位を送信する第２の送信手段とを備え、前記複数のノードのうちの少なくとも１つは、前記第１の送信手段による前記データの送信が終了した後、前記第１の変換単位情報記録手段に記録されている情報に対応するアドレス変換単位が前記第５の記憶手段に記憶されているかどうかを判定する判定手段と、前記判定手段による判定の結果、前記第５の記憶手段に記憶されていると判定されたアドレス変換単位を前記第２の送信手段に送信させるべき前記要求を前記通信路を介して行う送信要求手段と、をさらに備えるものとすることができる。

【００２１】前記第５の記憶手段に送信すべきデータのアドレス変換単位が記憶されているときは、その第５の記憶手段を有する前記複数のノードのうちのさらに他の少なくとも１つから前記複数のノードのうちの少なくとも１つにそのアドレス変換単位が送信されることとなるので、データの再送のために必要となる時間を短縮することができる。さらに、送信側となる前記複数のノードのうちの少なくとも１つにおける負担を軽減することができる。

【００２２】上記第１、第２の観点にかかるマルチプロセッサシステムにおいて、前記再送手段は、前記第１の変換単位情報記録手段に記録されている情報に対応するデータを前記第１のアドレス変換手段にアドレス変換させて前記第２の記憶手段から前記第１の記憶手段に書き込む手段を備え、該手段によって前記第１の記憶手段に書き込まれたデータを前記通信路を介して前記他のノードに再送するものとすることができる。

【００２３】上記目的を達成するため、本発明の第３の観点にかかるマルチプロセッサシステムにおけるデータ転送方法は、互いに通信路を介して接続され、それぞれ物理アドレス空間を提供する第１の記憶装置と、論理アドレス空間を提供する第２の記憶装置と、所定のアドレス変換単位毎に物理アドレスと論理アドレスとの間のアドレス変換を行うアドレス変換機構とを有する複数のノードから構成されるマルチプロセッサシステムにおけるデータ転送方法であって、前記複数のノードのうちの少なくとも１つから、該ノードの論理アドレス空間に存在するデータを前記通信路を介して他のノードに送信する送信ステップと、前記送信ステップで送信される前記データのうち前記第１の記憶装置が提供する前記物理アドレス空間に存在しなかったアドレス変換単位に関する情報を記録する第１の変換単位情報記録ステップと、前記送信ステップでの前記データの送信が終了した後、前記第１の変換単位情報記録ステップで記録した情報に対応

するアドレス変換単位のデータを前記複数のノードのうちの少なくとも１つから前記通信路を介して前記他のノードに再送する再送ステップと、前記送信ステップで前記複数のノードの少なくとも１つから送信されたデータを受信した前記他のノードが、該データを記憶すべき論理アドレスに対応する物理アドレスが前記第１の記憶装置に存在しないアドレス変換単位に関する情報を記録する第２の変換単位情報記録ステップとを含み、前記再送ステップは、前記第２の変換単位情報記録ステップで記録した情報に対応するアドレス変換単位のデータをさらに前記複数のノードのうちの少なくとも１つから前記通信路を介して前記他のノードに再送することを特徴とする。

【００２４】

【発明の実施の形態】以下、添付図面を参照して、本発明の実施の形態について説明する。

【００２５】［第１の実施の形態］図１は、この実施の形態にかかるマルチプロセッサシステムの機能構成を示す機能ブロック図である。図示するように、このマルチプロセッサシステムは、ネットワーク１０を介して互いに結合された複数のノード３－１～３－ｎからなる。

【００２６】ノード３－１～３－ｎは、同様の機能構成を有し、それぞれＣＰＵ（図示せず）、メモリ１３、二次記憶装置１８及び通信ハードウェア２７を備える。また、ノード３－１～３－ｎのそれぞれの上では、プロセス１１及びカーネル３６が動作する。このマルチプロセッサシステムで結合することが可能なノード３－１～３－ｎは、最大で２５６個とすることができる。

【００２７】メモリ１３は、ノード３－１～３－ｎのそれぞれの主記憶装置であり、物理アドレス空間を提供する。二次記憶装置１８は、磁気ディスク装置などによって構成され、ノード３－１～３－ｎのそれぞれに仮想アドレス空間を提供する。

【００２８】通信ハードウェア２７は、例えば、データの最大転送サイズを２５６ＫＢとし、ページサイズを４ＫＢとするページング方式で転送されるデータの仮想アドレスを物理アドレスに変換する機能を有する。通信ハードウェア２７は、また、割り込み発生部１７２、送信部１７１、受信部２７２及びリモートリード部５７１の各機能を有する。通信ハードウェア２７は、さらにプロセス１１からのデータの送受信の要求を書き込むための要求領域１７６と、要求されたデータの送受信が終了したことを示す情報を書き込むための確認領域１７５とを有する。要求領域１７６及び確認領域１７５は、仮想空間内にマップされている。

【００２９】割り込み発生部１７２は、アドレス変換に失敗したとき、すなわち送信或いは受信すべきデータについてページフォルトが発生したときに、ＣＰＵに割り込みを要求し、カーネル３６に割り込み処理をさせる。送信部１７１は、メモリ１３に書き込まれているデータ

を読み出し、パケットに分割してネットワーク１０に送出する。受信部２７１は、ネットワーク１０から受け取ったパケット中のデータをメモリ１３に書き込む。リモートリード部５７１は、ネットワーク１０から受け取ったリモートリード要求に応答して、カーネル３６の受信失敗ページ表６４１などを他のノードに読み込ませる機能を提供する。

【００３０】カーネル３６は、カーネル間通信や仮想記憶管理などの基本機能を提供するほか、割り込み処理３６２及び遠隔ページアクセス７６１の機能を提供する。カーネル３６は、また、仮想アドレス空間と物理アドレス空間との間でのアドレス変換を行うためのページテーブル４０と、データの送信時及び受信時にそれぞれページフォルトが発生したページに関する情報を記録するための送信失敗ページ表３４１と受信失敗ページ表６４１とを有する。

【００３１】割り込み処理３６２の機能は、ＣＰＵに割り込みが生じたときにその割り込み要因を特定し、特定した割り込み要因が送信部１７１或いは受信部２７１におけるアドレス変換の失敗（ページフォルト）である場合には、アドレス変換に失敗したページに関する情報を送信失敗ページ表３４１及び受信失敗ページ表６４１にそれぞれ記憶する。割り込み処理３６２では、送信失敗ページ表３４１或いは受信失敗ページ表６４１にアドレス変換に失敗したページに関する情報を記録した後、要求領域１７６にデータの送受信の再開要求を書き込むことにより、送信部１７１或いは受信部２７１の動作を再開させる。

【００３２】送信失敗ページ表３４１は、送信部１７１が送信するデータについてページフォルトが発生したときに、そのアドレス変換に失敗したページに関する情報を記録する。受信失敗ページ表６４１は、受信部２７１が受信するデータについてページフォルトが発生したときに、そのアドレス変換に失敗したページに関する情報を記録する。送信失敗ページ表３４１と受信失敗ページ表６４１とは、通信ハードウェア２７で転送可能なデータの最大サイズが決まっているため、図２に示すように、仮想アドレス＋ビット列でアドレス変換に失敗したページを記録する。

【００３３】これらの失敗ページ表では、図２に示すように、１ワード目には最初にアドレス変換に失敗したページのページ番号ＰＮＵＭが入れられる。２ワード目以降には、各ビットの値で対応するページがアドレス変換に失敗したかどうかを示し、例えば、２ワード目の第３１ビットの値が「１」であるならば、ＰＮＵＭ＋３１ページでアドレス変換に失敗したことを、３ワード目の第３１ビットの値が「１」であるならば、ＰＮＵＭ＋６３ページでアドレス変換に失敗したことを示すものである。なお、ＰＮＵＭの初期値は、「－１」とする。

【００３４】この実施の形態のマルチプロセッサシステムでは、データの送信側となるプロセス１１は、他の複数のノードのプロセスに対して同時にデータ転送を行うことはない。このため、ノード３－１～３－ｎのそれぞれにおいて、送信失敗ページ表３４１は１つだけ用意されている。一方、プロセス１１がデータの受信側となるときには、複数の他のノードのプロセスからデータを同時に受信することがあるので、受信失敗ページ表６４１は、送信側となるプロセスの数だけ用意される。なお、これらの送信失敗ページ表３４１及び受信失敗ページ表６４１の仮想空間上の配置を、図３に示す。

【００３５】遠隔ページアクセス７６１は、当該ノードとネットワーク１０を介して接続された他のノードにあるプロセスの仮想空間にアクセスし、アクセスしたページがページアウトされているときには、そのノードのカーネルにアクセスしたページをページインさせる。

【００３６】プロセス１１は、通信ハードウェア２７の要求領域１７６に送信要求を書き込むことによって、データ１２を他のノードのプロセスに送信する。プロセス１１は、通信ハードウェア２７の確認領域１７５の値を調べることによって、データ１２の送信が終了したことを確認する。プロセス１１は、他のノードから受信したデータを書き込むための受信バッファ２２を用意している。

【００３７】プロセス１１は、また、送信失敗ページ表３４１や受信失敗ページ表６４１を参照して、データの送受信時にアドレス変換の失敗があったかどうかを確認し、アドレス変換の失敗があったページについて再送処理を行う。なお、プロセス１１には、システム中でプロセスを一意に定められるように、「０」から「２５５」までのいずれかの識別子が割り付けられる。なお、データ１２及び受信バッファ２２は、仮想空間にあるもので、メモリ１３上に存在せずに、二次記憶装置１８にページアウトされていることがあり得る。

【００３８】以下、この実施の形態のマルチプロセッサシステムにおける動作について説明する。以下の説明において、ノード３－１をデータの送信側とし、ノード３－ｎをデータの受信側とする。ここで、ノード３－１のプロセス１１を送信側プロセス１１と呼び、ノード３－ｎのプロセス１１を受信側プロセス１１と呼ぶこととする。

【００３９】まず、ノード３－１とノード３－ｎとの間において、データ転送が開始される前に、ノード３－１とノード３－ｎとのそれぞれにおいて実行される準備処理について説明する。

【００４０】カーネル３６は、プロセス１１を生成したときに、生成したプロセス１１に対応するページテーブル４０と送信失敗ページ表３４１と受信失敗ページ表６４１を記憶する領域をメモリ１３上に確保し、これらをページアウトしないよう指定する。カーネル３６は、さらに送信失敗ページ表３４１と受信失敗ページ表６４１

とを、それぞれ仮想アドレスＳＮＤ＿ＦＴＢＬ＿ＡＤＤＲとＲＣＶ＿ＦＴＢＬ＿ＡＤＤＲにマップする。また、プロセス１１は、カーネル３６のマップ機能を用いて、要求領域１７６と確認領域１７５とを仮想空間内にマップする。

【００４１】次に、送信側ノード３－１から受信側ノード３－ｎへのデータ転送の大まかな流れについて、図４のフローチャートを参照して説明する。送信側プロセス１１は、まず、自プロセスの送信失敗ページ表３４１を初期化する。送信側プロセス１１は、さらに、通信ハードウェア２７の送信機能を利用して、受信側プロセス１１の受信失敗ページ表のうち送信側プロセス１１のものに対応する部分を初期化する。すなわち、送信側プロセス１１は、要求領域１７６に所定の送信要求を書き込んで、受信失敗ページ表６４１の初期値を受信側ノード３－ｎの該当領域に書き込む（ステップＳ１１）。なお、該当領域の計算は、数式１に従って行われる。

【数１】ＲＣＶ＿ＦＴＢＬ＿ＡＤＤＲ＋３ワード×プロセスの識別子

【００４２】送信側プロセス１１は、通信ハードウェア２７の要求領域１７６に所定の送信要求を書き込むことにより、通信ハードウェア２７にデータ１２の送信を要求する（ステップＳ１２）。

【００４３】送信側ノード３－１の送信部１７１は、要求領域１７６に書き込まれた送信要求を読み出すことによって、ネットワーク１０を介してのノード３－ｎへのデータ１２の送信を開始する。データ１２は、前述したようにパケットの形式で行われる。送信部１７１は、データ１２の途中でアドレス変換に失敗した場合には、データ１２の送信を停止して、ＣＰＵに対して割り込みを要求する（ステップＳ１３）。

【００４４】ステップＳ１３で割り込みが発生すると、送信側ノード３－１のカーネル３６は、割り込み処理を行い、割り込み処理の終了の後、送信部１７１にデータ１２の送信を続行させる（ステップＳ１４）。なお、この割り込み処理については、さらに詳しく後述する。

【００４５】受信側ノード３－ｎの受信部２７１は、ネットワーク１０から受け取ったデータを受信バッファ２２に書き込む。受信部２７１は、データの書き込みの途中でアドレス変換に失敗した場合には、データの受信を停止して、ＣＰＵに対して割り込みを要求する（ステップＳ１５）。

【００４６】ステップＳ１５で割り込みが発生すると、受信側ノード３－ｎのカーネル３６は、割り込み処理を行い、割り込み処理の終了の後、送信部２７１にデータ１２の送信を続行させる（ステップＳ１６）。なお、この割り込み処理については、さらに詳しく後述する。

【００４７】一方、送信側プロセス１１は、送信側ノード３－１の確認領域１７５を参照することによって、要求したデータ１２の送信が終了したことを確認する（ステップＳ１７）。

【００４８】送信側プロセス１１は、データ１２の送信が終了したことを確認すると、さらに送信側ノード３－１の送信失敗ページ表３４１と、受信側ノード３－ｎの受信失敗ページ表６４１とを参照して、アドレス変換の失敗があったかどうかをチェックする（ステップＳ１８）。ここで、送信側ノード３－１の送信部１７１でのアドレス変換の失敗は、通常のメモリアクセスで送信失敗ページ表３４１を読み、ＰＮＵＭの値が「－１」以外であれば発生していることがわかる。受信側ノード３－ｎの受信部２７１でのアドレス変換の失敗は、リモートリード部５７１を利用してリモートリードで受信側ノード３－ｎの受信失敗ページ表６４１を読み、ＰＮＵＭの値が「－１」以外であれば発生していることがわかる。

【００４９】ステップＳ１８で送信側ノード３－１の送信部１７１或いは受信側ノード３－ｎの受信部２７１のいずれかでアドレス変換の失敗があったことが確認された場合には、送信側プロセス１１は、必要に応じてデータ１２の再送処理を行う（ステップＳ１９）。なお、この再送処理については、さらに詳しく後述する。

【００５０】次に、送信側ノード３－１及び受信側ノード３－ｎのカーネル３６の割り込み処理３６２に機能が実行する割り込み処理（ステップＳ１４、Ｓ１６）について、図５のフローチャートを参照して詳しく説明する。割り込みが発生した場合、カーネル３６の割り込み処理４０の機能は、その割り込み要因について調べる（ステップＳ２１）。

【００５１】ステップＳ２１で調べた割り込み要因が送信部１７１におけるアドレス変換の失敗である場合には、この例では、送信側ノード３－１で生じた割り込みである。この場合は、データ１２の送信が停止されると共に、アドレス変換に失敗したページの仮想アドレスに関する情報が送信側ノード３－１の送信失敗ページ表３４１に記録される（ステップＳ２２）。このとき、送信側ノード３－１の割り込み処理部３６２は、当該ページのページイン要求を出し、メモリ１３にプリページングさせることができる。

【００５２】送信失敗ページ表３４１への記録が終了すると、データ１２の送信を再開させるための所定の送信要求が通信ハードウェア２７の要求領域１７６に書き込まれ、送信側ノード３－１の送信部１７１は、当該アドレス変換に失敗したページの次のページから、データ１２の送信を再開する。但し、アドレス変換に失敗したページが送信すべきデータ１２の最後のページであった場合には、データ１２の送信処理は終了する（ステップＳ２３）。

【００５３】ステップＳ２１で調べた割り込み要因が受信部２７１におけるアドレス変換の失敗である場合には、この例では、受信側ノード３－ｎで生じた割り込みである。この場合は、データ１２の受信が停止されると

共に、アドレス変換に失敗したページの仮想アドレスに関する情報が受信側ノード３－ｎの受信失敗ページ表６４１に記録される（ステップＳ２４）。このとき、このとき、受信側ノード３－ｎの割り込み処理部３６２は、当該ページのページイン要求を出し、メモリ１３にプリページングさせることができる。

【００５４】受信失敗ページ表６４１への記録が終了すると、受信側ノード３－ｎの受信部２７１は、当該アドレス変換に失敗したページを破棄し、次のページからのデータ１２を受信する。

【００５５】なお、ステップＳ２１で調べた割り込み要因がアドレス変換の失敗以外である時は、図５のフローチャートには示さないが、その割り込み要因に従った割り込み処理が行われる。

【００５６】次に、送信側プロセス１１が実行する再送処理（ステップＳ１９）について、図６（ａ）、（ｂ）のフローチャートを参照して詳しく説明する。送信側ノード３－１でのアドレス変換の失敗があった場合には、図６（ａ）に示すように、送信側プロセス１１は、送信側ノード３－１の送信側失敗ページ表３４１を参照して、アドレス変換に失敗したページの仮想アドレスを計算する（ステップＳ３１）。

【００５７】次に、送信側プロセス１１は、ステップＳ３１で計算した仮想アドレスにアクセスしてページフォルトを起こさせ、カーネル３６に当該アドレス変換に失敗したページを二次記憶装置１８からメモリ１３にページインさせる（ステップＳ３２）。そして、送信側プロセス１１は、そのページインされたページを再送する（ステップＳ３３）。

【００５８】一方、受信側ノードでのアドレス変換の失敗があった場合には、図６（ｂ）に示すように、送信側プロセス１１は、リモートリードで呼んだ受信側ノード３－ｎの受信失敗ページ表６４１を参照して、アドレス変換に失敗したページの受信側ノード３－ｎにおける仮想アドレスを計算する（ステップＳ４１）。

【００５９】次に、送信側プロセス１１は、カーネル３６の遠隔ページアクセス７６１の機能を利用して、受信側ノード３－ｎのステップＳ４１で計算した仮想アドレスにアクセスしてページフォルトを起こさせ、受信側ノード３－ｎのカーネル３６に当該アドレス変換に失敗したページを二次記憶装置１８からメモリ１３にページインさせる（ステップＳ４２）。そして、送信側プロセス１１は、当該受信側ノード３－ｎのメモリ１３に書き込めなかったデータを再送する。

【００６０】上記の処理で、送信側ノード３－１或いは受信側ノード３－２でアドレス変換に失敗したページが複数ある場合には、送信側プロセス１１は、そのすべてのページについてステップＳ３１、Ｓ３２、Ｓ４１、Ｓ４２の処理を行った上で、ステップＳ３３、Ｓ４３の処理を行う。

【００６１】なお、カーネル３６は、物理空間の空き容量が不足した場合には、ページアウトの処理を行うが、通信ハードウェア２７がページアウトされるページを参照している可能性もある。このため、カーネル３６がページアウトの処理を行い、ページテーブル４０に変更を加える場合には、通信ハードウェアの動作を一旦停止させて、通信ハードウェア２７の処理によってアドレス変換が行われないようにする。そして、ページテーブル４０の内容が変更された後に、通信ハードウェア２７の動作を再開させる。

【００６２】以下、図７を参照して、この実施の形態にかかるマルチプロセッサシステムの具体的な動作例について説明する。この例では、転送されるデータ１２は４Ｋバイトのページを４ページ含み、そのうち、１ページ目と４ページ目がメモリ１３上に存在し、２ページ目と３ページ目が二次記憶装置１８にページアウトされているものとする。

【００６３】プロセス１１がデータ１２を送信しようとした場合、１ページ目１２１の先頭アドレスは物理アドレス０ｘ１００００に変換される。プロセス１１は、その物理アドレスから１ページ分（０ｘ１０００バイト）のデータを送信部１７１に転送させる。このとき、ステップＳ１３のデータ送信では、アドレス変換の失敗はない。

【００６４】しかし、２ページ目１２２（０ｘ５０４０００～０ｘ５０７ＦＦＦ）はページアウトされているので、アドレス変換は失敗して割込み発生部１７２により割り込みが発生する。これにより、ステップＳ１３のデータ送信でアドレス変換の失敗があったことがカーネル３６に通知される。

【００６５】割り込み処理３６２ではまず、送信失敗ページ表３４１にこのページのページ番号０ｘ５０４を１ワード目に記録する。さらに、送信失敗ページ表３４１の２ワード目の第０ビットを「１」にする（ステップＳ２２）。このアドレス変換に失敗したページは、とばされて送信が進む（ステップＳ２３）。同様に、３ページ目もページアウトされているので送信失敗ページ表３４１の２ワード目の第１ビットも「１」とされる。

【００６６】送信側プロセス１１は、データ１２の送信終了後、失敗ページ表３４１を調べる（ステップＳ１８）。すると、送信側プロセス１１は、送信側で第２ページと第３ページにアドレス変換の失敗があったことがわかるので（ステップＳ３１）、これらのページにアクセスしてページフォルトを起こさせ、カーネル３６に第１ページと第２ページとを二次記憶装置１８からメモリ１３にページインをさせる（ステップＳ３２）。そして、送信側プロセス１１は、ページインが終った後で失敗した第２ページと第４ページのデータを再送する（ステップＳ３３）。

【００６７】以上説明したように、この実施の形態のマ

ルチプロセッサシステムでは、送信プロセス１１が送信失敗ページ表３４１或いは受信失敗ページ表６４１を参照してアドレス変換に失敗したページの有無をチェックした上でページインや再送処理を行うことができる。このため、送信プロセス１１が必要ないと判断したときは、ページインや再送処理を行わなくて済むので、無駄なページインの処理や再送処理を行わなくても済む。

【００６８】また、この実施の形態のマルチプロセッサシステムでは、転送されるデータについてアドレス変換に失敗したときに、ページインの要求だけを出しておき、それと並行してデータ転送を継続することが可能である。これにより、ページイン処理とデータ転送とをオーバーラップさせることができ、全体としてデータ転送を高速化することができる。

【００６９】［第２の実施の形態］この実施の形態では、データ転送を複数回行なった後で一回だけ失敗確認と再送をできるように失敗ページ表を変更することにより、失敗確認の回数を減らすことができるマルチプロセッサシステムについて説明する。

【００７０】この実施の形態のマルチプロセッサシステムの構成は、第１の実施の形態のもの（図１）とほぼ同一である。但し、この実施の形態のマルチプロセッサシステムでは、アドレス変換に失敗したページの仮想アドレスを記録するためのデータ構造として、第１の実施の形態の失敗ページ表の代わりに、失敗ページカウンタと失敗ページ表ポインタを使用する。

【００７１】図８は、失敗ページカウンタの構成を示す図である。失敗ページカウンタは、アドレス変換に失敗したページの数を保存するカウンタと、カーネルが失敗ページを記録するかどうかを指定するため書き込み許可フラグとから構成される。カウンタは３１ビットで構成され、データの転送を開始する前に送信プロセスによって「０」に初期化され、アドレス変換の失敗が起こる度にカーネル３６によって「１」ずつインクリメントされる。また、書き込み許可フラグは１ビットで構成されている。失敗ページを記録していないとき、すなわちデータ１２の送受信の終了後に失敗ページの確認の必要がない場合には、フラグは「０」に設定される。失敗ページを記録する場合には、フラグは「１に設定される。

【００７２】失敗ページ表ポインタは、アドレス変換に失敗したページの仮想アドレスを保持するための領域を指し示すポインタである。送信プロセス１１あるいは受信プロセス１１は、この仮想アドレスの記録領域をデータ転送の前に確保し、失敗ページ表ポインタに、その記録領域の先頭アドレスを設定する。

【００７３】このような失敗ページカウンタと失敗ページ表ポインタとからなるデータ構造は、プロセス毎にそれぞれ用意される。但し、第１の実施の形態の場合と異なり、受信用にはプロセス識別子の数だけこのようなデータ構造を用意すればよい。なお、プロセス１１の仮想空間における上記のデータ構造の配置の例を、図９に示す。

【００７４】以下、この実施の形態のマルチプロセッサの動作について説明する。カーネル３６は、プロセス１１を生成する際に、送信用および受信用の失敗ページカウンタ、書き込み許可フラグ、失敗ページ表ポインタの領域を確保し、決められた仮想アドレス領域にマップする。カーネル３６は、送信用および受信用の失敗ページカウンタ、書き込み許可フラグ、失敗ページ表ポインタの領域がページアウトされないように指定する。

【００７５】生成されたプロセス１１は、送信側、受信側ともに、データ転送に先だって、カーネル３６のメモリ割り当て機能を利用して、失敗ページ表の確保と設定を行なう。このメモリ領域は割り込み処理１７２によりアクセスされるので、カーネル３６は、失敗ページ表のメモリ領域がページアウトされないように指定する。例えば、Ｍａｃｈオペレーティングシステムの場合では、ｖｍ＿ｗｉｒｅシステムコールを使うことにより、失敗ページカウンタや失敗ページ表などのメモリ領域がページアウトされないように指定する。

【００７６】次に、プロセス１１は、このメモリ領域の先頭アドレスを送信用の失敗ページ表ポインタと受信用の失敗ページ表ポインタに設定する。プロセス１１が受信プロセスであり、データの送信を行なわない場合は送信用の失敗ページ表ポインタを設定しなくてもよい。また、受信用の失敗ページ表ポインタは通信する送信プロセスの分だけ設定する。

【００７７】送信側プロセス１１（ｐｉｄ＝ｎ）は送信を開始する前に、送信側の失敗ページカウンタを「０」に初期設定し、送信側の書き込み許可フラグを「１」に設定する。さらに、受信プロセスの仮想空間にある自分のプロセス識別子に対応する失敗ページカウンタのアドレス（ＦＴ＿ＢＡＳＥ＋２＊（ｎ＋１）＊４）に０ｘ８０００００００を転送する。

【００７８】そして、送信側プロセス１１は、通信ハードウェア２７の要求領域１７６にパラメータを書き込み、通信ハードウェア２７にデータ１２の転送を開始させる。

【００７９】送信側ノードの送信部１７１での処理においてアドレス変換の失敗が起こると、送信側ノードの通信ハードウェア２７の割り込み発生部１７２は、割り込みを発生する。送信側ノードの割り込み処理３６２の機能は、この割り込みを受け付けると、送信プロセス１１の識別子を通信ハードウェア２７の状態レジスタ（図示せず）からとりだす。割り込み処理３６２の機能は、送信側プロセス１１の仮想空間にある送信用の失敗ページカウンタの書き込み許可フラグを参照する。

【００８０】書き込み許可フラグの値が「１」であれば、割り込み処理３６２の機能は、まずその失敗ページカウンタの中のカウンタをインクリメントし、さらに、

失敗ページ表ポインタとカウンタの値から定まる場所（（先頭アドレス）＋（カウンタ）＊４）にアドレス変換に失敗したページの仮想アドレスを記録する。書き込み許可フラグの値が「０」であれば、割り込み処理３６２の機能では、失敗ページ表に対する操作はなにも行なわない。割り込み処理から復帰する際には、割り込み処理３６２の機能では、アドレス変換に失敗したページを飛ばすように指定して、送信側ノードの通信ハードウェア２７の送信部１７１のデータ転送を再開させる。

【００８１】一方、受信側ノードの受信部２７１での処理においてアドレス変換の失敗が起ると、受信側ノードの通信ハードウェア２７の割り込み発生部１７２は、割り込みを発生する。受信側ノードの割り込み処理３６２の機能は、この割り込みを受け付けると、送信プロセス１１の識別子と受信プロセス１１の識別子とを受信側ノードの通信ハードウェア２７の状態レジスタ（図示せず）から取り出す。割り込み処理３６２の機能は、受信プロセス１１の仮想空間にある受信用の失敗ページカウンタのうち送信プロセスに対応するものの書き込み許可フラグを参照する。

【００８２】書き込み許可フラグの値が「１」であれば、割り込み処理３６２の機能は、まず、その失敗ページカウンタの中のカウンタをインクリメントし、さらに、失敗ページ表ポインタとカウンタの値から定まる場所、例えば、失敗ページ表ポインタの値がＲＦＴｎ＿ＡＤＤＲ、カウンタの値がｋであったすると、ＲＦＴｎ＿ＡＤＤＲ＋４＊ｋのアドレスにアドレス変換に失敗したページのアドレスを書き込む。

【００８３】送信プロセス１１は、何回かデータ転送を行なうと、データ転送の途中でアドレス変換の失敗が起きたかどうかを確認する。送信プロセス１１は、送信側ノードの送信部１７１でのアドレス変換の失敗をチェックするために、自分の仮想空間にある送信用の失敗ページカウンタを読む。ここで、失敗ページカウンタ中のカウンタの値が「０」より大きければ失敗が起きたことになる。この場合は送信用の失敗ページ表ポインタの値から失敗ページ表の位置を特定し、そこからカウンタの分だけ失敗ページを読み出す。つまり、送信プロセス１１は、失敗ページポインタの値がＳＦＴ＿ＡＤＤＲ、カウンタの値がｋであった場合には、ＳＦＴ＿ＡＤＤＲのアドレスから４＊ｋバイト分読み出す。

【００８４】送信プロセス１１は、次に受信側での失敗をチェックするために、上述のリモートリード部５７１機能を使って受信プロセス１１の受信用の失敗ページカウンタ及び失敗ページ表ポインタのアドレスを同時に読む。この領域はページアウトされないのでリモートリードの途中でアドレス変換の失敗は起こらない。ここで、失敗ページカウンタ中のカウンタの値が「０」より大きければ失敗が起きたことになる。この場合は、カーネル３６が提供する遠隔ページアクセス７６１の機能を利用

して受信側ノードにある失敗ページ表をカウンタの分だけ読み出す。つまり、送信プロセス１１は、失敗ページポインタの値がＲＦＴｎ＿ＡＤＤＲ、カウンタの値がｋであった場合には、リモートリード機能によってＲＦＴｎ＿ＡＤＤＲのアドレスから４＊ｋバイト分読み出すこととする。

【００８５】なお、この実施の形態のマルチプロセッサシステムでは、送信プロセス１１は、あらかじめ確保された失敗ページ表を越すようなデータ転送を行なってはいけない。もしも、失敗ページ表を越すようなことが送信側ノードで起これば送信プロセス１１が、受信側で起これば受信プロセス１１がカーネル３６の機能によって異常終了させられる。このような異常終了を避けるためには、ユーザプログラムで最初に確保する失敗ページ表のサイズを大きくすればよい。

【００８６】以上説明したように、この実施の形態のマルチプロセッサシステムでは、複数のデータ転送に対する失敗確認と再送を一回で済ませることができるようになる。これにより、別々のアドレスにある小量のデータを送るような場合に、失敗確認の回数を減らすことができるので、全体としてデータ転送に要する時間を第１の実施の形態の場合よりもさらに短縮することができる。

【００８７】［第３の実施の形態］この実施の形態では、例えば、転送されるデータ１２の最後の１ビットをデータの受信完了を示すフラグとして用いている。この実施の形態にかかるマルチプロセッサシステムの構成は、第１の実施の形態のものとほぼ同一であるが、通信ハードウェア２７の割り込み発生部１７２の機能として、次の２つの機能が追加されている。

【００８８】すなわち、割り込み発生部１７２は、送信部１７１がデータ１２を送信するときに、送信すべきデータのうちの最後のパケットとなったときに割り込みを発生し、受信部２７１がネットワーク１０からデータ１２を受信するときに、受信すべきデータのうちの最後のパケットとなったときに割り込みを発生する。

【００８９】以下、この実施の形態のマルチプロセッサシステムにおける動作について説明する。カーネル３６の割り込み処理３６２の機能は、第１の実施の形態で説明した割り込み処理に加えて、次のような割り込み処理を行う。なお、この割り込み処理以外の動作は、第１の実施の形態における動作と同一である。

【００９０】図１０は、この実施の形態において、送信側ノード３－１と受信側ノード３－ｎのカーネル３６の割り込み処理３６２の機能が実行する割り込み処理を示すフローチャートである。割り込みが発生した場合、カーネル３６の割り込み処理４０の機能は、その割り込み要因について調べる（ステップＳ２１）。

【００９１】ステップＳ２１で調べた割り込み要因が送信すべきデータのうちの最後のパケットとなったものである場合は、この例では、送信側ノード３－１で生じた

割り込みである。この場合、割り込み処理３６２の機能は、まず、送信側ノード３－１の送信失敗ページ表３４１を調べる（ステップＳ５１）。そして、送信失敗ページ表３４１を調べた結果、送信部１７１の処理でいずれかのページにアドレス変換の失敗があったかどうかを判定する（ステップＳ５２）。

【００９２】ステップＳ５２でいずれかのページにおいて、いずれかのページにアドレス変換の失敗があったと判定したときは、送信側ノード３－１の割り込み処理３６２の機能は、送信失敗ページ表３４１に最後のページについて（実際にはアドレス変換の失敗がなくても）アドレス変換の失敗があったものとして送信側ノード３－１の送信失敗ページ表３４１に記録する。そして、最後のパケットの送信を停止させて（ステップＳ５３）、この割り込み処理を終了する。

【００９３】一方、ステップＳ５２でいずれのページにもアドレス変換の失敗がなかったと判定したときは、そのまま割り込み処理を終了する。この場合、最後のパケットは送信されることとなる。

【００９４】ステップＳ２１で調べた割り込み要因が受信すべきデータのうちの最後のパケットとなったものである場合は、この例では、受信側ノード３－ｎで生じた割り込みである。この場合、割り込み処理３６２の機能は、まず、受信側ノード３－１の受信失敗ページ表６４１を調べる（ステップＳ５４）。そして、受信失敗ページ表６４１を調べた結果、受信部２７１の処理でいずれかのページにアドレス変換の失敗があったかどうかを判定する（ステップＳ５５）。

【００９５】ステップＳ５５でいずれかのページにおいて、いずれかのページにアドレス変換の失敗があったと判定したときは、受信側ノード３－ｎの割り込み処理３６２の機能は、受信失敗ページ表６４１に最後のページについて（実際にはアドレス変換の失敗がなくても）アドレス変換の失敗があったものとして受信側ノード３－ｎの送信失敗ページ表６４１に記録する。そして、受信した最後のパケットを破棄させて（ステップＳ５６）、この割り込み処理を終了する。

【００９６】一方、ステップＳ５５でいずれのページにもアドレス変換の失敗がなかったと判定したときは、そのまま割り込み処理を終了する。この場合、受信した最後のパケットは、受信バッファ２２（メモリ１３上にある）に書き込まれる。

【００９７】以上のように割り込み処理を加えたことによって、アドレス変換に失敗したページを再送する場合には、最終ページも再送されることとなり、受信側ノード３－ｎでは、データの受信完了を示すフラグを含む最後のパケット（最終ページの最終パケット）を１番最後に受信することとなる。

【００９８】以上説明したように、この実施の形態のマルチプロセッサシステムでは、受信側ノード３－ｎは最後のパケットを必ず最後に受信することとなるので、例えば、データ中の最後の１ビットを受信完了を示すフラグとして用いるプログラムでも、正しく動作させることが可能となる。

【００９９】［第４の実施の形態］図１１は、この実施の形態にかかるマルチプロセッサシステムの機能構成を示す機能ブロック図である。このマルチプロセッサシステムは、第１の実施の形態のマルチプロセッサシステム（図１）とほぼ同様の機能構成を有するが、カーネル４６が他のノードのカーネルからの要求を受けて、二次記憶装置１８に存在するデータを送信先ノードに送信する遠隔データ送信１０６２の機能を有する点が異なる。

【０１００】以下、この実施の形態にかかるマルチプロセッサシステムにおける動作について説明する。この実施の形態のマルチプロセッサシステムの動作は、第１の実施の形態のものとほぼ同一であるが、ステップＳ１９の再送処理のみが異なる。また、ステップＳ１９の再送処理においても、受信側ノード４－ｎでアドレス変換の失敗があった場合の処理は、第１の実施の形態のもの（図６（ｂ））と同一である。

【０１０１】図１２は、送信側ノード４－１でアドレス変換の失敗があった場合に、送信側プロセス１１が実行する再送処理を示すフローチャートである。処理が開始すると、送信側プロセス１１は、送信側ノード４－１の送信側失敗ページ表３４１を参照して、アドレス変換に失敗したページの仮想アドレスを計算する（ステップＳ３１）。

【０１０２】次に、送信側プロセス１１は、ステップＳ３１で計算した仮想アドレスが示すページが遠隔ディスク（送信側ノード４－１以外にある二次記憶装置１８）にあるかどうかを判定する（ステップＳ３４）。

【０１０３】ステップＳ３４で当該ページが遠隔ディスクにないと判定したときは、送信側プロセス１１は、ステップＳ３１で計算した仮想アドレスにアクセスしてページフォルトを起こさせ、カーネル４６に当該アドレス変換に失敗したページを二次記憶装置１８からメモリ１３にページインさせる（ステップＳ３２）。そして、送信側プロセス１１は、そのページインされたページを再送する（ステップＳ３３）。

【０１０４】ステップＳ３４で当該ページが遠隔ディスクにあると判定したときは、送信側プロセス１１は、カーネル４６の遠隔データ送信要求１０６２の機能を利用して、当該ノードのカーネル３１に対して遠隔データ送信要求を送る（ステップＳ３５）。そして、送信側ノード４－１における処理を終了する。

【０１０５】次に、この遠隔データ送信要求を受信したノード（以下、遠隔ノードという）における処理を、図１３のフローチャートを参照して説明する。遠隔ノードの遠隔データ送信１０６１の機能は、遠隔データ送信要求を受け取ると、その要求中に含まれている情報を元

に、遠隔ノードの二次記憶装置１８から読み出すべきページを決定する（ステップＳ６１）。

【０１０６】次に、遠隔データ送信１０６１の機能は、二次記憶装置１８からステップＳ６１で決定したページを読み出し（ステップＳ６２）、受信側ノード４－ｎにそのページを　送信る（ステップＳ６３）。最後に、遠隔データ送信１０６１の機能は、送信が終了したことを要求元の送信側プロセス１１に遠隔ノードのカーネル４６が提供するプロセッサ間通信機能を使って通知する（ステップＳ６４）。以上により、一連の再送処理を終了する。

【０１０７】以上説明したように、この実施の形態のマルチプロセッサシステムでは、遠隔ノードから直接データを送信するので、遠隔ノードから送信側のノードへのデータ転送を減らすことができ、データの再送の時間を短縮することができる。さらに、送信側ノードの負担を減らすこともできる。

【０１０８】［実施の形態の変形］上記の第１の実施の形態では、プロセス１１が受信バッファ２２のページインさせるためにカーネルが提供する遠隔ページアクセス機能７６１を利用したが、受信側ノードにページアクセス専用のスレッドを設け、そのスレッドを使ってページインさせてもよい。

【０１０９】このページアクセススレッドは要求メッセージを受けとると、受信バッファ２２の指定された仮想アドレスに直接アクセスすることによりページフォルトを起こし、カーネル３６にページインさせる。この要求メッセージは受信プロセスの識別子とページインさせたい仮想アドレスのリストを含む。なお、この要求メッセージは従来のカーネルが提供するのと同様のシステムコールを用いた通信機能を用いて送ることができる。

【０１１０】この方法を用いることにより、カーネルが遠隔ページアクセス機能７６１を持たない場合でも、第１の実施の形態と同様のマルチプロセッサシステムを構築することができる。

【０１１１】上記の第３の実施の形態では、最後のパケットを送信／受信する時に割り込みを起こし、その割り込み処理の中で最後のパケットを送信／受信しないようにする処理を行なっているが、最後のパケットを送信／受信する時の割り込みの代わりに、最後のパケットあるいは最後のページを送信／受信しないで捨てるような機能が通信ハードウェアにある場合も、同様のことが実現できる。

【０１１２】このような通信ハードウェアを使用した場合、最初にアドレス変換の失敗が起きた時の割り込み処理で、送信／受信失敗ページ表には最後のページも失敗したものとして記録し、さらに、通信ハードウェアに対して最後のパケット（あるいはページ）の送信／受信をしないで捨てるように要求する。これにより、最後のパケット（あるいはページ）は送信／受信されず、後で送

信プロセスが失敗確認した時には最後のページもアドレス変換に失敗したように見える。これにより、上記の第３の実施の形態と同様の効果を得ることができる。

【０１１３】上記の第１～第４の実施の形態では、ページング方式の仮想記憶をサポートするものに本発明を適用した場合について説明したが、セグメント方式やセグメントページング方式の仮想記憶をサポートする疎結合型のマルチプロセッサシステムにも本発明を適用することは可能である。

【０１１４】

【発明の効果】以上説明したように、本発明によれば、ノード間でのデータ転送を高速化することができる。また、不必要なアドレス変換や再送処理を行わなくてもよい。

【図面の簡単な説明】

【図１】本発明の第１の実施の形態にかかるマルチプロセッサシステムの機能構成を示す機能ブロック図である。

【図２】本発明の第１の実施の形態における失敗ページ表（送信失敗ページ表及び受信失敗ページ表）の構成を示す図である。

【図３】本発明の第１の実施の形態における失敗ページ表の仮想空間上の割付を示す図である。

【図４】本発明の第１の実施の形態にかかるマルチプロセッサシステムにおける動作を示すフローチャートである。

【図５】本発明の第１の実施の形態においてカーネルが実行する割り込み処理を示すフローチャートである。

【図６】本発明の第１の実施の形態においてプロセスが実行する再送処理を示すフローチャートであり、（ａ）は送信側で失敗があった場合の処理を、（ｂ）は受信側で失敗があった場合の処理を示す。

【図７】本発明の第１の実施の形態にかかるマルチプロセッサシステムの動作例を説明する図である。

【図８】本発明の第２の実施の形態に適用されるページカウンタの構成を示す図である。

【図９】本発明の第２の実施の形態における失敗ページ表の仮想空間上の割付を示す図である。

【図１０】本発明の第３の実施の形態においてカーネルが実行する割り込み処理を示すフローチャートである。

【図１１】本発明の第４の実施の形態にかかるマルチプロセッサシステムの機能構成を示す機能ブロック図である。

【図１２】本発明の第４の実施の形態においてプロセスが実行する再送処理を示すフローチャートであり、送信側で失敗があった場合の処理を示す。

【図１３】本発明の第４の実施の形態においてプロセスが実行する遠隔データ送信処理を示すフローチャートである。

【符号の説明】

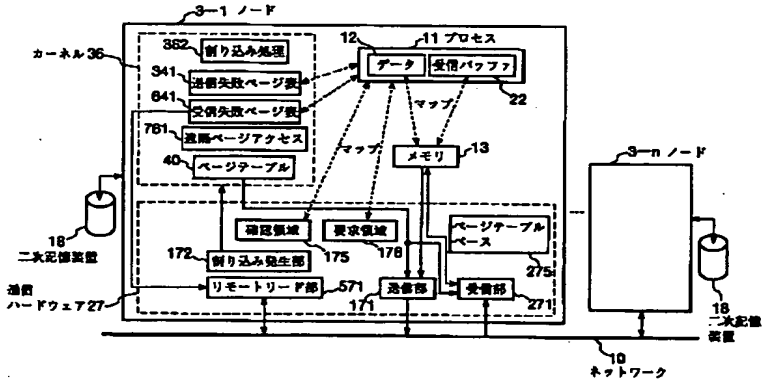
25

３－１、３－ｎ、４－１、４－ｎ　ノード
１０　ネットワーク
１１　プロセス
１２　データ
１３　メモリ
１８　二次記憶装置
２２　受信バッファ
２７　通信ハードウェア
３６、４６　カーネル
４０　ページテーブル
１７１　送信部
１７２　割り込み発生部

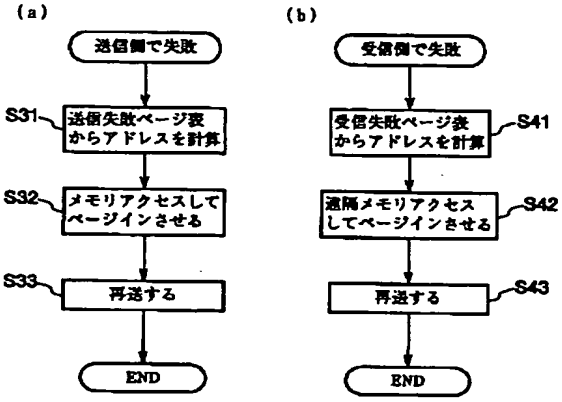
26

１７５　確認領域
１７６　要求領域
２７１　受信部
２７５　ページテーブルベース
３４１　送信失敗ページ表
３６２　割り込み処理
５７１　リモートリード部
６４１　受信失敗ページ表
７６１　遠隔ページアクセス
１０６１　遠隔データ送信
１０６２　遠隔データ送信要求

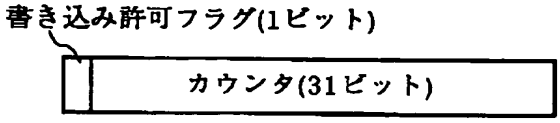
【図１】



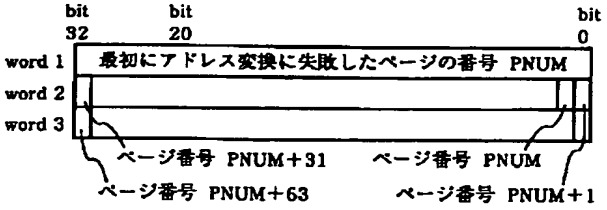
【図６】



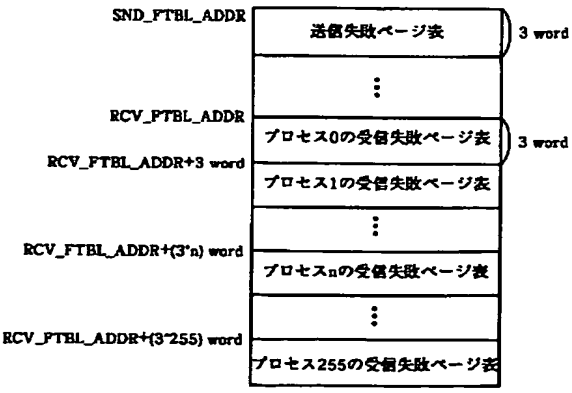
【図８】



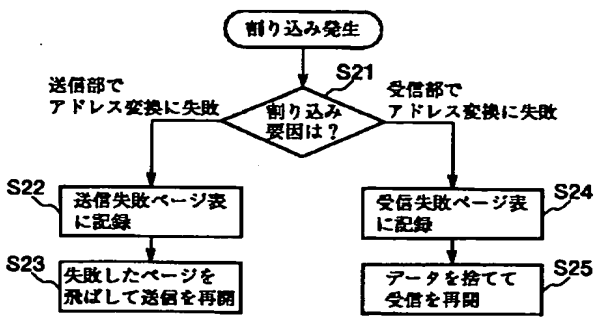
【図２】



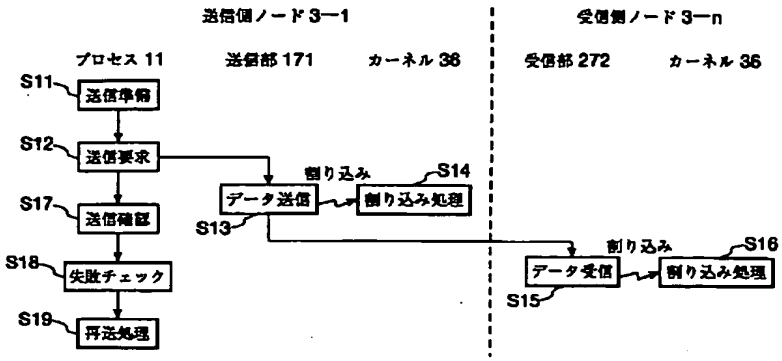
【図３】



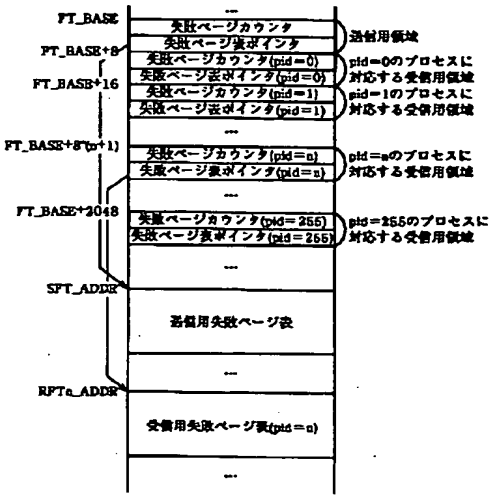
【図５】



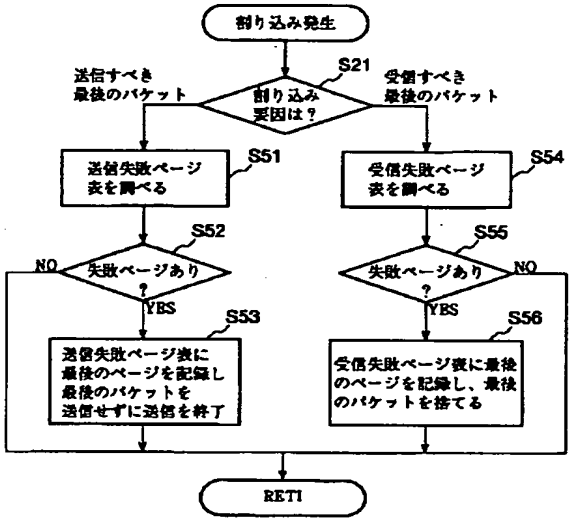
【図４】



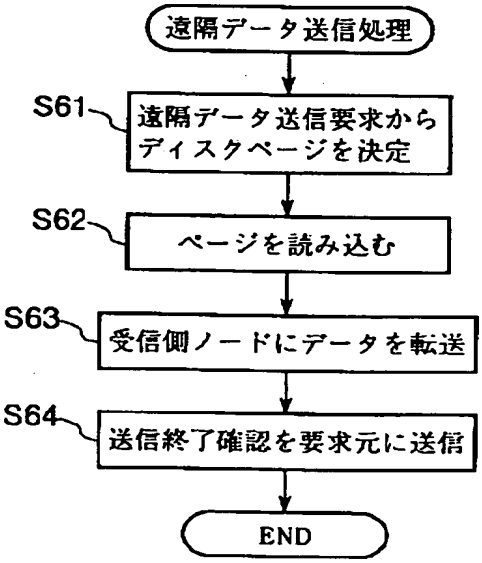
【図９】



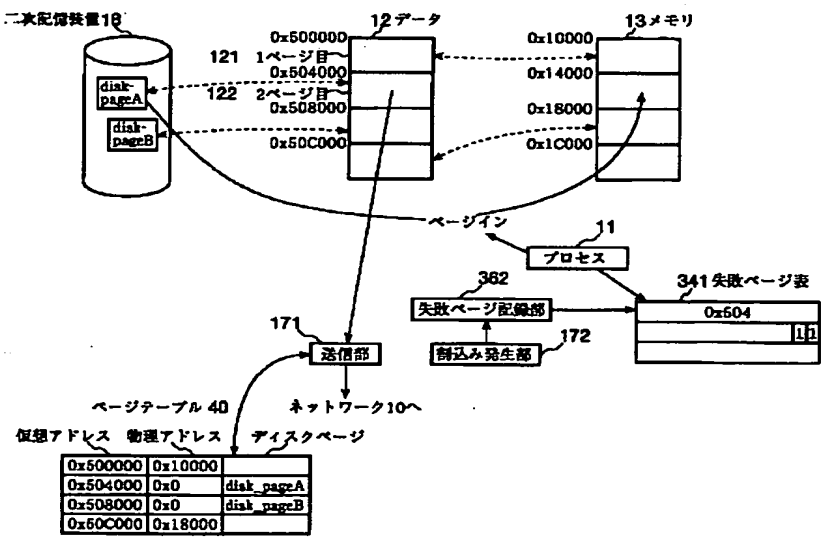
【図１０】



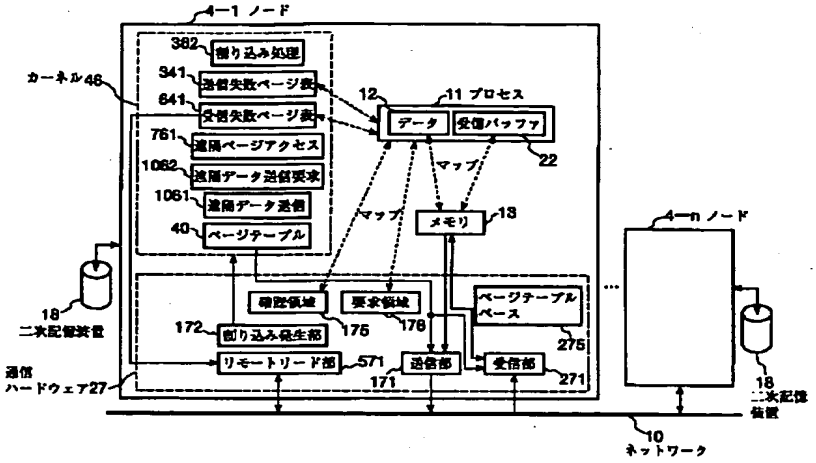
【図１３】



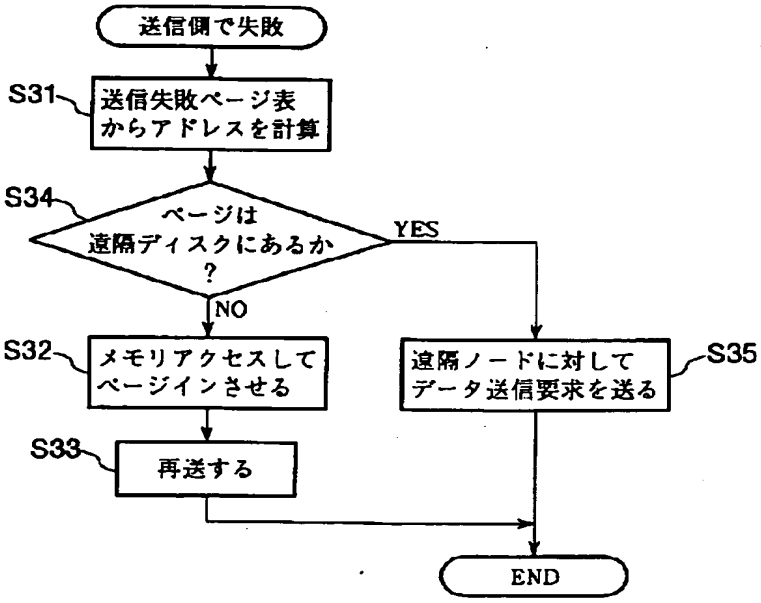
【図７】



【図１１】


4−1 ノード
382 割り込み処理
341 送信失敗ページ表
641 受信失敗ページ表
761 遠隔ページアクセス
1062 遠隔データ送信要求
1061 遠隔データ送信
40 ページテーブル
カーネル46
12
11 プロセス
データ
受信バッファ
22
マップ
メモリ
13
4−n ノード
18 二次記憶装置
172 割り込み発生部
175
176
ページテーブルベース
275
リモートリード部
571
171
送信部
受信部
271
通信ハードウェア27
10 ネットワーク

【図１２】


送信側で失敗
S31
送信失敗ページ表からアドレスを計算
S34
ページは遠隔ディスクにあるか？
YES
NO
S32
メモリアクセスしてページインさせる
S33
再送する
S35
遠隔ノードに対してデータ送信要求を送る
END

フロントページの続き

(56)参考文献　特開　昭57－162164（ＪＰ，Ａ）
特開　平７－64846（ＪＰ，Ａ）
特開　平７－271739（ＪＰ，Ａ）
特開　平９－212474（ＪＰ，Ａ）
特開　平８－305667（ＪＰ，Ａ）
特開　平７－262151（ＪＰ，Ａ）
手塚宏史、外４名，「ピンダウンキャッシュを用いたユーザレベルゼロコピー通信」，情報処理学会研究報告Ｖｏｌ．97　Ｎｏ．76（97－ＡＲＣ－125），1997．08．22，ｐ．167－172
石川裕，「コモディティハードウェアを用いた並列処理技術」，情報処理第39巻第８号，1998．08．15，ｐ．784－791

(58)調査した分野(Int.Cl.[7]，ＤＢ名)
G06F　15/16 - 15/177
G06F　12/08 - 12/12
G06F　13/00
G06F　13/38 - 13/42
ＪＩＣＳＴファイル（ＪＯＩＳ）